Beauty, a road to the truth

Kuipers, T.A.F.

Published in:
Synthese

THEO A. F. KUIPERS

BEAUTY, A ROAD TO THE TRUTH[1]

ABSTRACT. In this article I give a naturalistic-cum-formal analysis of the relation between beauty, empirical success, and truth. The analysis is based on the one hand on a hypothetical variant of the so-called 'mere-exposure effect' which has been more or less established in experimental psychology regarding exposure-affect relationships in general and aesthetic appreciation in particular (Zajonc 1968; Temme 1983; Bornstein 1989; Ye 2000). On the other hand it is based on the formal theory of truthlikeness and truth approximation as presented in my *From Instrumentalism to Constructive Realism* (2000). The analysis supports the findings of James McAllister in his beautiful *Beauty and Revolution in Science* (1996), by explaining and justifying them. First, scientists are essentially right in regarding aesthetic criteria useful for empirical progress and even for truth approximation, provided they conceive of them as less hard than empirical criteria. Second, the aesthetic criteria of the time, the 'aesthetic canon', may well be based on 'aesthetic induction' regarding nonempirical features of paradigms of successful theories which scientists have come to appreciate as beautiful. Third, aesthetic criteria can play a crucial, schismatic role in scientific revolutions. Since they may well be wrong, they may, in the hands of aesthetic conservatives, retard empirical progress and hence truth approximation, but this does not happen in the hands of aesthetically flexible, 'revolutionary' scientists.

> *We may find totally opposite things beautiful: a simple mathematical principle as well as a series of unrepeatable complex contingencies. It is a matter of psychology.* (Stephen Jay Gould, translated passage from (Kayzer 2000, 30)).

1. INTRODUCTION

Elucidating the concepts of the True, the Good, and the Beautiful and elucidating their relations, is frequently called the classical task of philosophy. In my opinion, this task should be carried out as much as possible in agreement with scientific insights and findings. Moreover, if possible, it should be demystifying, and, if necessary, even disenchanting. The topic of this paper is the relation between truth and beauty. Among scientists the intuition is widespread that there is a strong bond between them. Dirac is the classical representative of this intuition. Many other examples of natural scientists expressing this opinion could be given. For example, in a recent interview series on Dutch television, entitled 'Concerning Beauty and Consolation',[2] the physicist Steven Weinberg and the biologist Stephen Jay

Synthese **131:** 291–328, 2002.

Gould emphasised that considerations of beauty play an important role in their appreciation of theories and findings. The best philosopher's response seems to me to be, OK, this intuition of scientists should be taken seriously, but since it is not self-evident, to say the least, an explanation and justification of it should be given. However, to be honest, before reading McAllister (1996), I always thought, like many other philosophers of science,[3] that there must be some kind of misunderstanding among scientists. It not only seemed implausible that aesthetic appreciation has anything to do with empirical success, let alone with truth or truth approximation, but even a relation between nonempirical features frequently mentioned as examples of aesthetic features, such as symmetry and simplicity, on the one hand and truth approximation on the other seemed rather implausible on the basis of my own approach to truth approximation.

1.1. *Some Examples and Preliminary Considerations*

So let us be open-minded and try to explain and justify the fact that purported truths are frequently deemed beautiful[4] and that this is why the beauty of an idea, of a potential truth, is considered to be an indication of its soundness. Expressions such as 'the splendour of truth' and 'the simplicity of truth' support this. However, as James McAllister has shown in his very inspiring book *Beauty and Revolution in Science* (1996) (for brief expositions, see also McAllister 1998, 1999), our aesthetic judgements are subject to change. We are not only inclined to find the heliocentric worldview of Copernicus more beautiful, because it is simpler, than the geocentric view of Ptolemy, but we are also inclined to find Kepler's elliptic planetary orbits at least as beautiful as Corpenicus's circular orbits. However, ellipses are undoubtedly more complicated than circles, and this is precisely the reason why they were found less beautiful, if not problematically ugly, at the time. Moreover, I would like to add, aesthetic criteria not only change with time within a discipline, but may also differ greatly between disciplines. For example, after expressing in the interview series mentioned above his agreement with Weinberg about the importance of beauty considerations, Gould hastens to stress that his criteria for beauty totally differ from those of Weinberg. Whereas Weinberg mentions 'inevitability' of desired consequences as his dominant aesthetic criterion – as exemplified by Einstein's theory, which, in contrast to Newton's theory, made the inverse square in the law of gravitation inevitable – Gould stresses that, besides diversity, unrepeatable contingencies and irregularities are the sources of his ultimate aesthetic satisfaction.[5] Ironically enough, Weinberg not only refers to Bach's music in general and his two-part Inventionen in particular as showing a similar kind of inevitability, but also

mentions the gravedigger scene in Shakespeare's Hamlet as a surprising intermezzo in a logical sequence of events, which, according to Weinberg, illustrates the fact that in the arts there are even higher aesthetic phenomena than in science.[6]

Having observed the variation of aesthetic criteria, McAllister's main claims are as follows. First, scientists normally use aesthetic criteria in addition to empirical criteria for theory evaluation. Second, and most importantly, the aesthetic criteria of the time, the 'aesthetic canon', is based on 'aesthetic induction' regarding nonempirical features of paradigms of empirically successful theories which scientists have come to appreciate as beautiful. Third, aesthetic criteria can play a crucial, schismatic role in scientific revolutions. Since they may well be wrong, they may, in the hands of aesthetic conservatives, retard empirical progress and even truth approximation, but this does not occur in the hands of aesthetically flexible, 'revolutionary' scientists.

1.2. *Outline of the Paper*

In this article I present an analysis of the relation between the truth and the beauty of scientific theories that, in the end, elaborates and supports McAllister's claims. Like McAllister, I will concentrate on nonempirical aesthetic features, that is, features with aesthetic value but without empirical content, although empirical features may also be aesthetically valued. In Section 2 I will first argue, in the spirit of naturalized epistemology, that the phenomenon of aesthetic induction may be a variant of the so-called 'mere-exposure effect', and then decompose the notion into aesthetic induction proper and a related cognitive (meta-) induction. Together they lead to correlations between nonempirical features which are found beautiful on the one hand and empirically successful theories on the other. Such correlations will be called 'beauty-success correlations'. The corresponding received or 'canonical' aesthetic features are nonempirical features that have acquired (positive) aesthetic value and (empirical success related) inductive support. Moreover, this makes it plausible to explicate the notion of an 'aesthetic feature' as an aesthetically (positively) valued nonempirical (objective) feature.

In the rest of the paper I will argue that the co-production of the two types of induction is functional for empirical progress and even for truth approximation as far as the cognitive meta-induction is reliable. For this purpose I will present, in Section 3, the basic definition of greater truthlikeness, which, roughly, refers to a theory allowing more desirable and fewer undesirable possibilities than another. Subsequently I will argue that this definition can be rephrased as, again put roughly, a theory having fewer

undesirable and more desirable features than another. In Section 4 the difference between empirical and nonempirical features will be explicated, followed by an elaboration of how claims to truth approximation can be judged in terms of empirical criteria, especially in terms of observed (and therefore desirable) possibilities, representing instantial successes, on the one hand and established desirable observational features, representing explanatory successes, on the other. In Section 5 the title question will come under discussion when I examine the relative importance of empirical and aesthetic considerations. I will do this by comparing the importance of relevant differences between two theories in the light of the hypothesis that one is closer to the truth than the other: a difference in explanatory success, a difference in instantial success, and a difference in 'aesthetic success', that is, (not) having a received aesthetic feature. One of the outcomes will be that an aesthetic success can be just as good a signpost to the truth as an extra case of explanatory success, albeit in a more modest degree. The relevant difference is that the justified desirability of such an explanatory success can be more reliably established than that of an aesthetic feature, which is why the latter should be approached with more care. In Section 6 this comparative analysis will enable us to point out the heuristic-methodological use of aesthetic features in seven different problem situations, amongst which is a typical 'revolutionary' one. Some suggestions for further research will be given in Section 7, followed by several conclusions in Section 8.

1.3. *Some Limitations and Specifications*

As I said, the formal analysis is based on the basic theory of truthlikeness and truth approximation by empirical progress. Some limitations and specifications should be mentioned beforehand.

First, in view of the fact that the basic theory, and for that matter the refined version too, has, up until now, mainly been restricted to the natural sciences, this study will be too.

Second, 'the truth' will be understood as the strongest, that is to say the most informative, true theory of what is physically (chemically, biologically) possible, within the scope of a domain and a vocabulary that have been chosen beforehand. Thus 'the truth' is always conceived as the product of language and reality; therefore there can be many truths. Because these truths are connected in many ways (one truth can, for instance, be reducible to another), this is not an extremely relativistic position, but one that has a realistic tendency. Neither is it an extreme metaphysical, essentialist position, for the language in question is a human construction, not an ideal language that is supposed to be somehow inextricably con-

nected to the natural world. The name 'constructive realism' covers these two aspects. Beyond this, it is important to note that the definition of 'the truth', though it depends on the idea of a 'true theory', is not circular. I presuppose a definition of a true theory, namely as true for all (bio-)physical possibilities. Here 'true for a physical possibility' is defined in accordance with Tarski's truth definition.

Third, our truth approximation claims regarding aesthetic features are, at least in this paper, restricted to a certain formal type of aesthetic features. More precisely, the 'underlying' objective nonempirical features of aesthetic features, and objective (nonempirical and empirical) features of theories in general, will be restricted to a certain formal type. A feature of a theory is called 'distributed' when it corresponds to an objective property of all (formal representations of) the conceptual possibilities admitted by the theory. Note first that aesthetic features of theories are not supposed to be associated with (the set of) its real world instances, but with the corresponding (set of) conceptual possibilities. However, it may well be that the aesthetic appreciation concerns a non-formal type of representation of certain conceptual possibilities. The famous Feynman diagrams in quantum electrodynamics provide an example. But also in such a case, it is assumed that there is in addition a formal, i.e., logico-mathematical, representation of the conceptual possibilities, such that the aesthetic feature is co-extensional with an objective property of the relevant formal conceptual possibilities. The corresponding distributed feature is called the objective feature underlying the aesthetic feature. Aesthetic features of which the objective nature cannot be explicated in the suggested distributed way fall outside the scope of my truth approximation claims, and demand further investigation. However, it should be stressed that some standard aesthetic features are of the distributed type. Regarding simplicity, for example, it is important to note that the members of the set of conceptual possibilities satisfying a simple formula all share the property to 'fit' in this simple formula. Regarding symmetry, representing a kind of order, we may note that a theory is frequently called symmetric because all its possibilities show a definite symmetry. For example, all admitted orbits may have a certain symmetrical shape. Regarding inevitability and its opposite contingency (see below), it is also plausible to assume that at least certain types of both properties can be localised within conceptual possibilities.

Fourth, and finally, in this article I will not elaborate on the historical examples of truth approximation and aesthetic considerations. For the former I refer to Kuipers (2000, Ch. 10 and 11), for the latter I refer to McAllister's book and articles. But now and then I will refer to features

that are widely considered beautiful, like simplicity and order in the form of symmetry.

In summary, the answer to the question of the title will be: yes, beauty can be a road to the truth, namely as far as the truth is beautiful in the specific sense that the truth has distributed features that we have come to experience as beautiful due to (a variant of) the mere-exposure effect. It will become clear that this is a nontrivial answer. Though it may be a disenchanting one, this conclusion has heuristic-methodological use for truth approximation, provided that the aesthetic criteria, in comparison to the empirical criteria, are handled with great caution. The answer can very well be considered as an explication of what McAllister can meaningfully have in mind when he speaks of the relation between beauty, empirical success, and truth in terms of aesthetic induction.

2. AESTHETIC INDUCTION AND EXPOSURE EFFECTS

McAllister (1996, also 1998) introduces the notion of 'aesthetic induction'. It refers to the phenomenon that scientists tend aesthetically more and more to appreciate recurring nonempirical, objective features of successful theories. In this section I want to disentangle this phenomenon. In order to do so, I will first relate it to experimental psychological studies of emotive *e*ffects of repeated exposure to certain stimuli, that is, studies of exposure-*a*ffect relations in general and aesthetic appreciation in particular.

2.1. *Exposure Effects and Aesthetic Appreciation*

McAllister (1998) gives an architectural illustration of the phenomenon in question. To colour photographs of the Washington Monument and the Eiffel Tower he adds the following telling caption:

> Architectural aesthetics, as embodied in the Washington Monument (left) and the Eiffel Tower (right), can change with the discovery of a material's structural utility – much like the aesthetic appeal of a scientific theory, which seems to grow with every empirical success. The Washington Monument (1884), a white marble obelisk, pays homage to the aesthetic canons of ancient Egypt. In contrast, the Eiffel Tower, which was built a mere five years afterward, is an iconoclastic cast-iron structure that displayed aesthetic properties unprecedented in its day. Cast-iron architecture eventually attained wide use, and a broad aesthetic appeal, as its utility was discovered.

That we grow to find something beautiful after acceptance of it is widely acknowledged where the arts are concerned. Marjoleine de Vos (1999) even goes as far as saying about canonised poetry:

...but the not-beautiful, if accepted, in point of fact always changes into the beautiful. Time makes everything beautiful. And habituation helps.

In popular music the plug-effect is well known. By repeating the same song or video-clip time and again many people tend to notice by themselves that they like it more and more. This is evidently exploited for commercial purposes by popular broadcasting channels. Similarly, it is well known that the number of people appreciating the atonal music of Schönberg and their degree of appreciation initially were very low, but that both the number and the degree increased considerably with time. See Mull (1957) for early experimental evidence on this.

All these phenomena seem to be related to the so-called 'mere-exposure effect' which has been studied in experimental psychology (Zajonc 1968; Temme 1983; Bornstein 1989; Ye 2000). Various experiments, for example with music, paintings, drawings, photos, Chinese characters, and advertisements, illustrate the fact that an increasing number of presentations of the same item tends to increase the aesthetic or, at least, affective appreciation of that item. However, frequently one observes not only first a phase of monotone increasing aesthetic appreciation with the number of confrontations, but also a second phase of decreasing appreciation, together producing the so-called 'inverse U' shape. The most plausible explanation is of course that in the end people get bored. After first having the excitement of recognising more and more in combination with seeing or hearing more and more, one may become so used to a piece of art that both types of excitement fade away. The Bolero of Ravel might be an illustration. This explanation is known as the two-factor model of stimulus habituation and satiation. Although many comparative studies with varying experimental conditions prompting or retarding the switch have been done (see e.g., Bornstein 1989; Ye and Van Raaij 1997; Ye 2000), two conditions that are particularly interesting for our purposes, viz., successive variation of the same stimulus and introducing some kind of reinforcement, have not been studied, as far as I know. In music one might think of retardation of the switch point by introducing some variation, e.g., in a musical theme, or by presenting different performances of the same piece of music. This would be interesting because theory revision may frequently be considered as variation on a theme. Since 'mere exposure' is by definition unreinforced, experiments with various kinds of reinforcement would deviate from the paradigmatic type of research in this area. Reinforcement would of course be interesting because (increasing) empirical success evidently is a type of reinforcement in theory revision, in the same way as McAllister suggests in the caption quoted above that the utility of cast-iron architecture reinforces its ascribed aesthetic value. However this may be, it is not only established

that the mere-exposure effect occurs under certain conditions and within certain limits, but also seems plausible that variation and reinforcement will retard satiation. The latter hypothetical variant of the mere-exposure effect will be called the (postulated) qualified-exposure effect.

Let me illustrate the mere-exposure effect by an informal, but nevertheless risky experiment, just initiated by my own experience. The television series mentioned above comprised 25 interviews with various celebrities in the sciences, the humanities, and the arts. Each installment of the programme started with the same intro of 1.5 minute, namely, a tango danced by an elegant and charming older couple accompanied by light effects. Halfway through the series, I presented the main ideas of the present paper in Amsterdam for an audience of 30 people. After showing the intro by video, but before giving my expectations, I ascertained that about half the audience had never or almost never seen the programme before. Only 40% of these said that they found the video particularly beautiful. In contrast, of the other half, that is, of those who had seen the programme at least a number of times, 80% found the video particularly beautiful. In agreement with my own experience, most of the latter confirmed that their appreciation for the intro had gradually increased. It should be noted that there was no variation in the intro and, more importantly, something like intellectual reinforcement was not evidently the cause of the increasing appreciation of the series. For, like most commentators in the media, one had become more and more critical about the obsessive, non-stimulating way in which Kayzer interviewed his impressive guests.

The mere-exposure effect and the suggested qualified-exposure effect call for general explanations. There must be some kind of (related) psychophysical mechanisms producing them. Using Tinbergen's four well-known questions about biological behavioural patterns, there is reason to investigate the causal structure of the mechanisms, the direct or proximate functions of the effects, the ontological development of the mechanisms, and the phylogenetic, hence evolutionary, development, explaining the distal functions of the effects, that is, the indirect functions that more directly serve reproduction and survival. Regarding the mere-exposure effect itself, the two factors postulated by the two-factor theory, recall, habituation and satiation, certainly are the plausible psychological point of departure for searching for a general causal mechanism behind them. In the last decade it has been argued that there is a link with implicit learning and memory (Bornstein 1994; Seamon et al. 1995). Regarding the evolutionary background, Bornstein (1989) concludes with some general speculations. I will deal with only some of these and related questions, as far as aesthetic appreciation in science is concerned.

2.2. *Aesthetic Induction in Science*

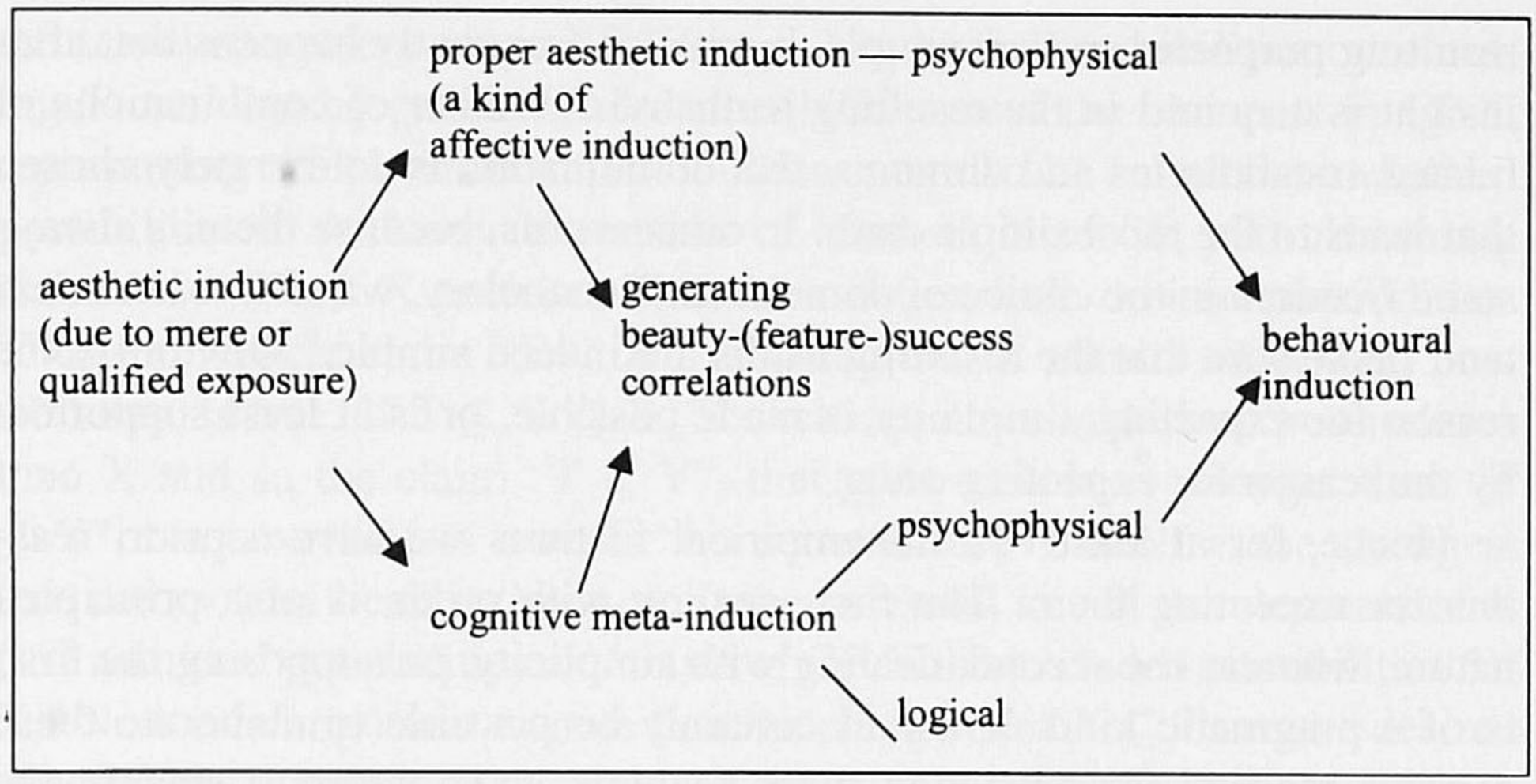
It is evident that McAllister's notion of 'aesthetic induction' can be seen as a reinforcement variant of the mere-exposure effect. More specifically, McAllister claims that aesthetic induction is triggered by empirical success, i.e., in psychological terms, empirical success functions as a kind of reinforcement. If the number of empirically successful theories with a certain nonempirical feature increases the aesthetic appreciation of that feature increases. Similarly, if increasingly many empirically successful revisions of a theory have a constant nonempirical feature, that feature becomes aesthetically more and more appreciated. This phenomenon naturally leads to McAllister's idea of an 'aesthetic canon' of received aesthetic features in a certain phase of a discipline that may be replaced by a different one after a scientific revolution. And, I like to add, the canon may be different for coexisting research programmes within one discipline and for different disciplines, depending on the specific nonempirical features of successful theories in the respective programmes and fields.

From now on we should sharply distinguish between aesthetic features and merely nonempirical, but objectively, or at least intersubjectively, determinable features of theories. The latter will briefly be called 'nonempirical' features. Aesthetic features – more precisely, the canonical ones – are here conceived as nonempirical features which (certain) scientists (have come to) find beautiful, that is, to which they ascribe aesthetic value. Typical examples of frequently mentioned aesthetic features are symmetry, simplicity, visualizability, and inevitability.[7] From the present point of view it is perfectly possible that other scientists also call their opposites, viz., asymmetry, complexity/diversity, abstractness, contingency, aesthetic, especially in other periods or other disciplines.

The distinction between nonempirical features and their possible ascribed aesthetic value makes it possible to disentangle the kinds of induction that may be involved when a nonempirical feature accompanies empirical success. In fact, one can distinguish at least two kinds of induction, one of an emotive or affective and one of a cognitive nature.[8] The underlying idea of the suggested qualified-exposure effect in a scientific context is that there is a psychophysical mechanism that first arouses some kind of aesthetic appreciation for a nonempirical feature of a new, empirically successful theory, which, if positive, subsequently increases due to repetition of this feature in other successful theories. If the early response is negative, the idea is of course that this first diminishes and then switches to positive appreciation, which subsequently increases. But, for simplicity, I assume from now on that the start is positive. Note that the arousal as such is analogous to the generation of charge or current

in electromagnetic induction, but note also that the latter does not have the increasing character which the former is supposed to have. Be this as it may, it is plausible to speak here of *proper aesthetic* (or more generally, emotive or *affective*) *induction*, by definition of an increasing nature. At the same time, the first co-occurrence of a nonempirical feature and success may arouse the cognitive expectation that it will also accompany the next successful theory and this expectation may be strengthened by subsequent co-occurrences of the feature and empirical success. This will be called *cognitive meta-induction*, where 'meta' refers to the fact that we are dealing with features of theories, rather than *object-induction* dealing with features of objects in the natural world. Whereas affective induction has only one (psychophysical) side, cognitive (meta-) induction has two sides: a psychophysical side and a logical side. The psychophysical side of cognitive induction in general is usually assumed to have some, albeit weak, logical justification: inductive expectations and generalisations have some kind of support, called inductive support, and even on the meta-level of theories, there may be some truth in them of the following type: in order to be empirically successful, theories (in a certain area and perhaps even within a certain research programme) may well require a certain nonempirical feature.

The result of both types of induction is that correlations gradually grow between success and beauty, mediated by features, called beauty-feature-success or, simply, beauty-success correlations, in the following sense: recurring nonempirical features of successful theories come to be found beautiful. According to this diagnosis, canonical aesthetic features are nonempirical features that have acquired both (positive) aesthetic value and (empirical success-related) inductive support. The analysis suggests the following related important questions with respect to the correlation-producing mechanism. (1) Is the correlation mechanism functional in evolutionary perspective? The answer is certainly 'yes', if the answer to the second question is so. (2) Is the correlation mechanism functional for empirical progress? The answer to this question is certainly positive as far as (the logical side of) cognitive meta-induction is reliable. The reason is that, in addition to cognitive meta-induction and perhaps other cognitive reasons, the aesthetic appreciation resulting from proper aesthetic induction further strengthens the search for theories having the relevant nonempirical features, which might be called *behavioural induction*. Of course, it is far from evident that, or to what extent, cognitive meta-induction is reliable. This is a general question, to which I need for present purposes only a very cautious positive answer. In the present, naturalistic approach, this positive answer may be based on the standard view of naturalized epistemology: in



Scheme 1. A decomposition of 'aesthetic induction' in science.

view of the widespread inductive habits of humans and higher animals, these habits apparently have survival value. That is, they must bring us on average onto the right track, or at least more frequently than other types of systematic expectation formation, including random formation. For if they did not, other learning-from-experience strategies would have become dominant.

Scheme 1 summarizes the resulting disentanglement of McAllister's notion of aesthetic induction in science.

As far as some standard examples of aesthetic features in physics are concerned, viz. order and simplicity, it is possible to give, in addition, a priori reasons to expect them to be features of the truth, and hence, according to the formal analysis which is to follow, features of empirically successful theories. The first reason concerns the order in (the truth about) reality. Physics presupposes, to a large extent, the so-called principle of the uniformity of nature: not everything is physically possible and what is, does not depend on place and time. Assuming a certain vocabulary we try to grasp by means of our theories what is physically possible and, in view of the principle of uniformity, we assume this to have a certain order. If everything were physically possible, then there would be no, or at least less, order. Although the principle cannot be proved, it would be hard to explain the results of physics, unless the principle were true: scientific truths, i.e., laws of nature, are possible only when there is a certain order or system in reality.

Alongside order, simplicity is a nonempirical feature frequently mentioned by physicists as an aesthetic feature. This is at least partly a consequence of the fact that science is the work of humans. A vocabulary and domain cannot only have been chosen accidentally such that the

resulting purported truth is simple, but it also frequently happens that after insight is acquired in the resulting truths of a number of combinations of related vocabularies and domains, that combination is deliberately chosen that leads to the most simple truth. In other words, because there is always some freedom in the choice of domain and vocabulary, we can, to some extent, make sure that the resulting truths are indeed simple.[9] Obviously the reason for expecting simplicity is made possible, or is at least supported, by the reason for expecting order.

Hence, for at least two nonempirical features we have a priori reasons for expecting them. The first, dealing with order, is of a principled nature, whereas the second, dealing with simplicity, presupposing the first, is of a pragmatic kind. It would certainly be possible to elaborate these a priori reasons for expecting these features and further to specify the relation between them. Moreover, for other nonempirical features, other a priori reasons to expect them, in physics or in other natural sciences, can probably be given. However, this is not important for my present purposes. As far as such a priori reasons are valid, they make sure that the processes of cognitive meta-induction and proper aesthetic induction will start and continue to work, but such reasons are of course not necessary for this purpose. It may well be a contingent fact, also supported by the principle of the uniformity of nature, that empirically successful theories, and the corresponding truths, have certain nonempirical features. Both types of induction will operate in these cases just as well.

In so far as cognitive meta-induction is reliable, a positive answer can be given to the resulting explication of the title question of this paper: (3) Is the beauty-success correlation mechanism functional for truth approximation? The conditionally positive answer to this main question requires first an exposition of the basic theory of truth approximation.

3. TRUTHLIKENESS

The starting point of the idea of truthlikeness[10] is a vocabulary and a domain. A conceptual possibility is a situation or state of affairs that is describable in the vocabulary, and therefore conceivable. Let CP be the set of all conceptual possibilities describable in terms of the vocabulary. A theory will be associated with a subset of CP. A basic assumption is that the representation of the chosen domain in terms of the vocabulary results in a subset of CP containing the physical possibilities. We can identify this usually unknown subset with the truth T for reasons that will become clear shortly. For the sake of convenience I will assume that we can somehow characterise T in terms of the vocabulary. The aim of theory formation is

the actual characterisation of T. Hence, the physical possibilities constituting T can also be called desired possibilities, and the elements in CP-T, representing the physical impossibilities, can also be called the undesired possibilities.[11] A theory X consists of a subset X of CP, with the strong claim "X = T". If X encloses T, X does not exclude desired possibilities. Thus the weaker claim "T $\subseteq$ X", meaning that X admits all desired possibilities, is true. If T $\subseteq$ Y $\subseteq$ X, Y excludes more undesired possibilities than X and so the claim "T $\subseteq$ Y", that goes with it, is stronger than "T $\subseteq$ X", but nevertheless true. In this sense theory T itself is the strongest true theory, and I call it *the truth*. It seems useful to call the elements of X (its) admitted possibilities and those of CP-X the excluded possibilities (of X). Now it is important to note that the elements of X $\subseteq$ T are the desired possibilities admitted by X, and X-T consists of the undesired possibilities admitted by X. In Figure 1 all four resulting categories are depicted.

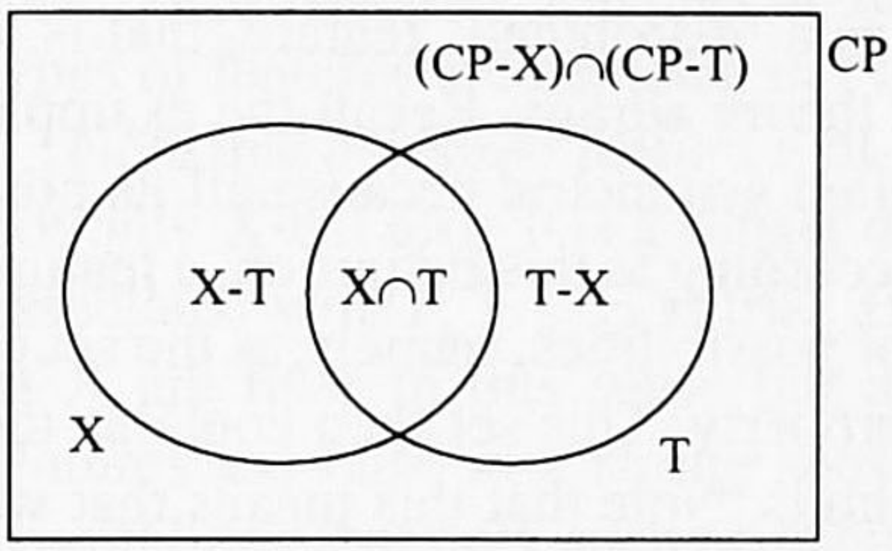


Figure 1. Four categories of possibilities.

This brings us directly to the basic definition of (equal or greater) truthlikeness:

DEFINITIONS

Y is at least as close to T as X (or: Y resembles T as much as X) iff

(DP) all desired possibilities admitted by X are also admitted by Y

(UP) all undesired possibilities admitted by Y are also admitted by X

Y is (two-sided) closer to T than X (or: Y resembles T more than X) iff

(DP) & (DP+) Y admits extra desired possibilities

(UP) & (UP+) X admits extra undesired possibilities

Figure 2 indicates which sets must be empty (clause (DP) and (UP): vertical and horizontal shading, respectively) and which sets have to be

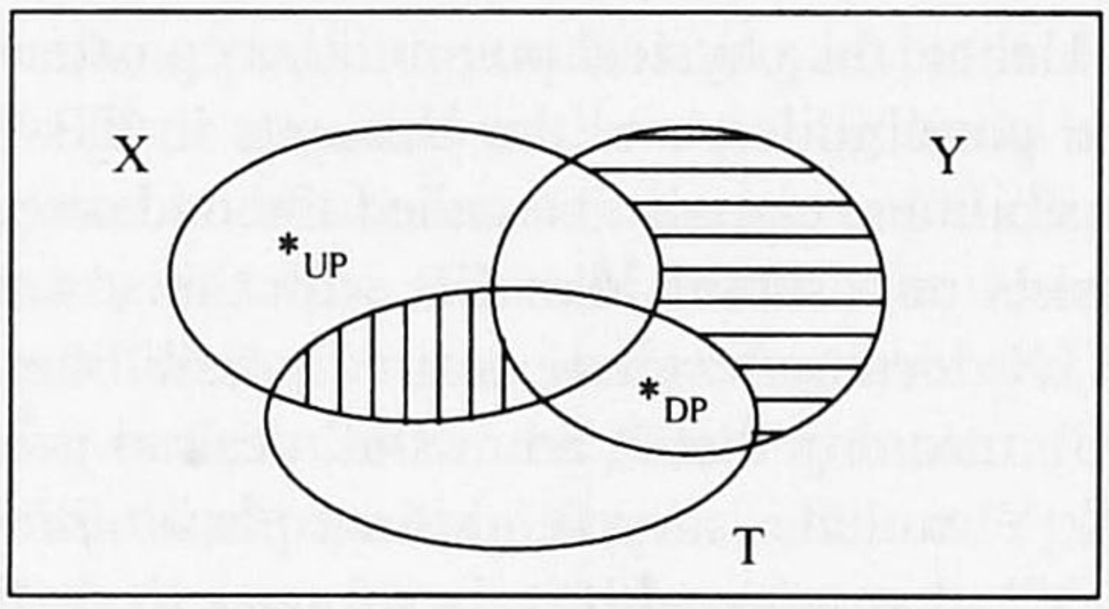


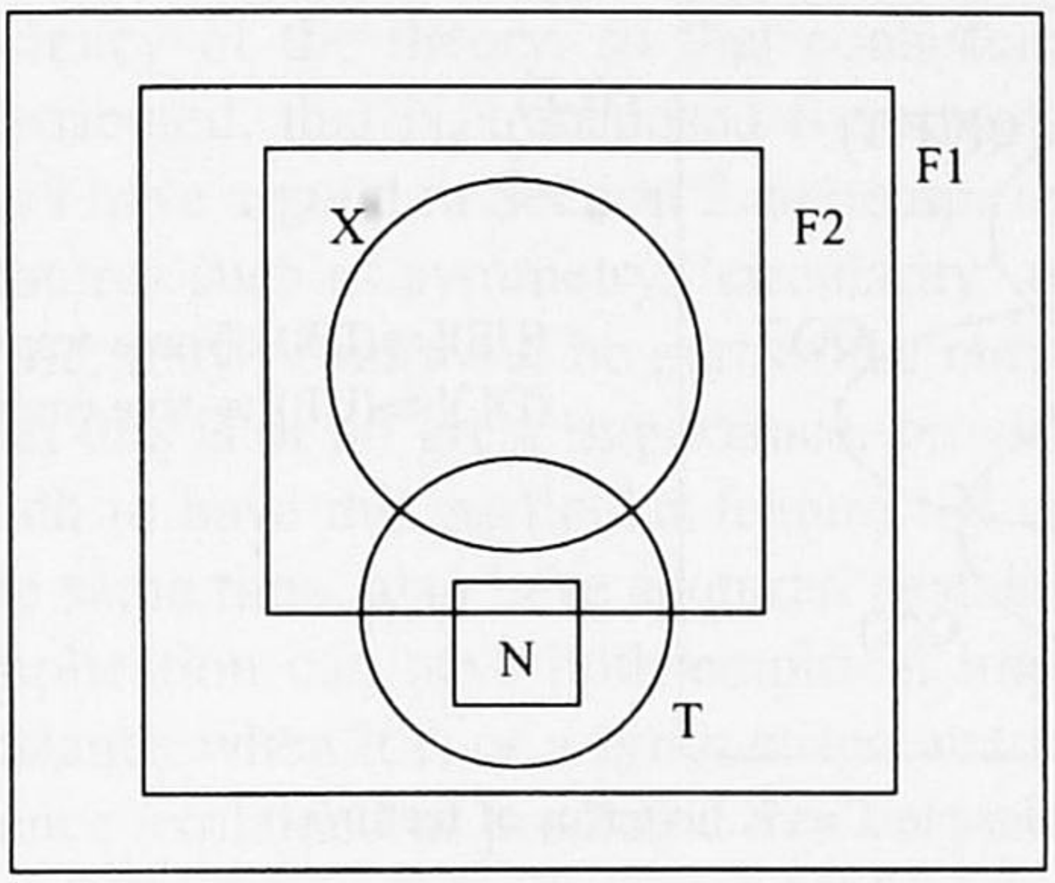
Figure 2. Y is closer to the truth T than X.

non-empty (clause (DP+) and (UP+): area *DP and area *UP non-empty, respectively) in the case that Y is closer to the truth than X.[12]

It is of great importance to our main question that the definitions can be reformulated in terms of desirable and undesirable features of a theory. The starting point consists of properties of possibilities. As announced, a feature of a theory will be understood as a 'distributed' feature, that is, a property of all the possibilities that the theory admits. Recall the example of symmetry. A theory is frequently called symmetric because all its possibilities show a definite symmetry.[13] According to this definition, a feature of a theory can be represented as a set of possibilities, namely as the set of all possibilities that have the relevant property. This set then contains the set of all possibilities that the theory admits. Note that this means that we could say that a feature of a theory excludes (exactly) all possibilities that do not have that property.

By way of preview we can easily see already now that nonempirical features of theories may have something to do with truth approximation by theories, for both, that is, theories and their features, are reconstrued as sets of possibilities. At first sight it may in particular have been surprising that aesthetic properties of theories can be cast in terms of sets of possibilities, since they seem to relate to form or structure. However, as has been indicated, as soon as it is a common characteristic of all the possibilities of a theory, which is frequently the case, it can be represented as a feature in our distributed sense.

By now it is obvious how we can formulate explicit definitions of desired, undesired, and remaining features in terms of the (logical) exclusion of desired and undesired possibilities: desired features are features that include all desired possibilities or, equivalently, that exclude only undesired possibilities; undesired features are features that include all undesired possibilities or, equivalently, that exclude only desired possibilities. All remaining features, as far as they can be represented as a subset of CP, exclude desired and undesired possibilities; that is, they do not include either all desired possibilities or all undesired ones. These are features



F1: a desired feature of X
F2: a neutral feature of X
F3 = CP − N: an undesired feature of X

Figure 3. Three types of features.

about which we can be neutral, for which reason I call them neutral features. However, they will play no role in the following analysis.[14] The three types of features are depicted in Figure 3.

Note that a desired feature F of X is a true feature of X in the sense that not only X but also T is a subset of F, that is, the weak claim that may be associated with F, $T \subseteq F$, is true. However, not only all undesired features of X are false in this sense but also all neutrai features. The undesired features are false in a strong sense: they not only exclude some desired possibilities, but only such possibilities.

The following theorems can now easily be proved (see note 16):

EQUIVALENCE THESES

Y is at least as close to T as X iff

(UF) all undesired features of Y are also features of X
(equivalent to (DP))

(DF) all desired features of X are also features of Y
(equivalent to (UP))

Y is two-sidedly closer to T than is X iff

(UF) & (UF+) X has extra undesired features (equivalent to (DP+))

(DF) & (DF+) Y has extra desired features (equivalent to (UP+))

In Figure 4, 'at least as close to the truth' is depicted in terms of features. The rectangle now represents the 'universe' of all possibly relevant,

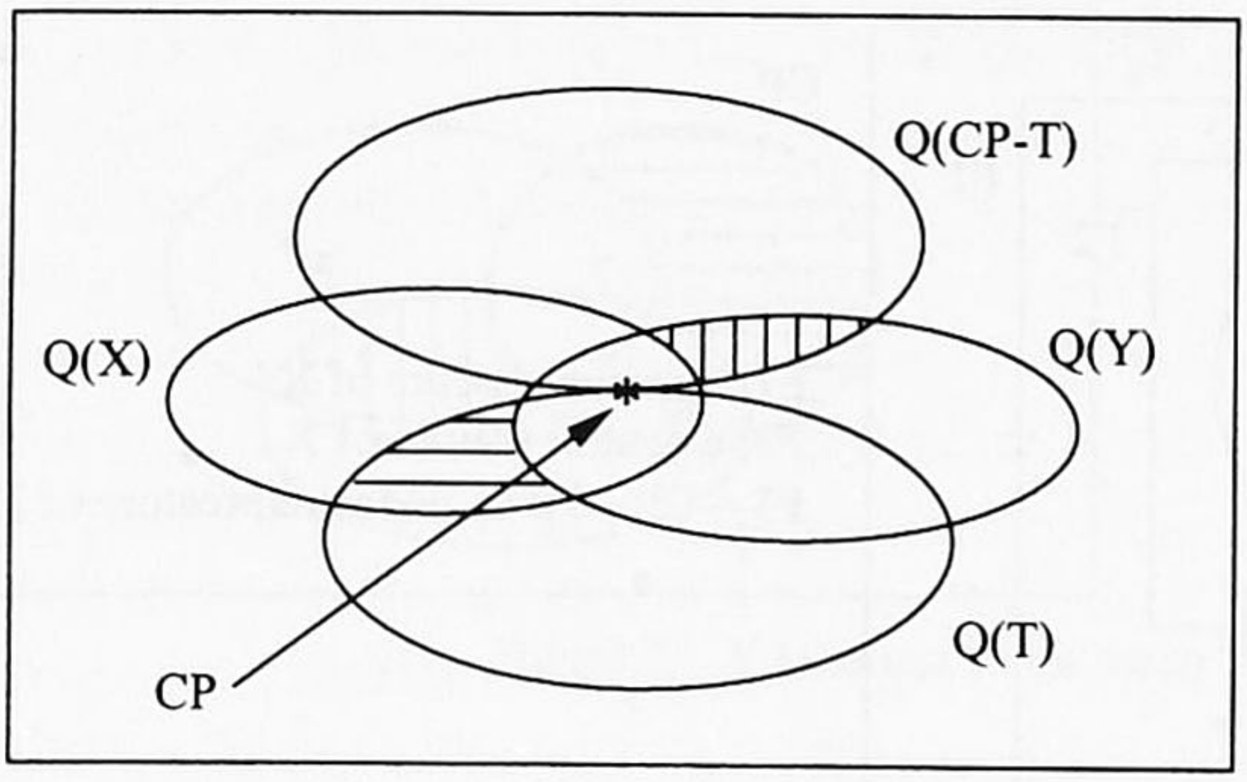


Figure 4. Y is at least as close to T as X, in terms of features.

distributed, features. Q(X) and Q(Y) represent the set of features of X and Y, Q(T) represents the set of desired features (features of T) and Q(CP-T) represents the set of undesired features (the features of CP-T). Note that, Q(T) and Q(CP-T) have exactly one element in common, namely the tautology, which can be represented by CP.[15]

Notice the strong analogy between the logical form of (DP) and (DF) and between that of (DP+) and (DF+). The same goes for the logical form of (UP) and (UF), and that of (UP+) and (UF+). The equivalencies as stated in the theorem, though, correspond in the reverse way: (DP) and (UF) are equivalent, as are (DP+) and (UF+), (UP) and (DF), (UP+) and (DF+). All this is not at all surprising, for undesired features could be defined in terms of desired possibilities and vice versa. Therefore it is in principle possible to reproduce the proof of the theses informally, clause by corresponding clause.[16]

On the basis of the equivalencies, it follows that the two principal definitions can also be given in a mixed or 'dual' form, in terms of desired possibilities and desired features: (DP) and (DF) for at least as close to the truth, with the addition of (DP+) and (DF+) for (two-sidedly) closer to the truth. The above provides us with all ingredients of the basic theory of truthlikeness that are needed to answer our main question.

4. TRUTH APPROXIMATION WITH THE AID OF EMPIRICAL CRITERIA

In the preceding section I have discussed objective features of theories in general. Of course there are different kinds of features. An obvious classification is the division into empirical and nonempirical features. The nonempirical features are dividable into logical and non-logical or conceptual ones. The best-known and most important logical feature is consistency. Notice that every permitted possibility points to the con-

sistency of the theory, so that consistency, like many other features, is distributed, that is, transposed from possibilities to the level of theories. As I have argued in Section 2, nonempirical, more specifically, conceptual features such as symmetry, 'circularity' or satisfying a conservation principle, may or may not be part of the current aesthetic canon. We will see that this is of no great importance, provided and as long as we expect the truth to have this particular feature. Of course, empirical features can, at the same time, also have acquired aesthetic value. A certain observational implication can have both empirical importance and aesthetic value, for instance when it is of a symmetrical nature or when it can be derived, and hence explained or predicted in a very nice way. But in cases such as this we will take the line that the empirical feature is primary and the aesthetic value secondary. Here I will focus on nonempirical aesthetic features, such that their importance and role can be examined in the purest form. But first we will study empirical criteria for theory evaluation.

There are two main categories of empirical criteria of a theory, in accordance with the dual design above. I have already mentioned the question whether or not the theory implies a certain established observational law or regularity, that, if so, can be explained or predicted by the theory. The implication of an observational law can thus be conceived as an established desired observational feature of the theory. Observational laws are of course established by 'object induction' on properties recurring in repeated experiments. Instead of speaking of implication or explanation and/or prediction of the theory, in what follows I will simply speak of explanation of such laws. Besides the 'explanation criterion' there is the 'instantial criterion', viz., the admission or exclusion of an observed possibility, that is, the result of a particular experiment being an example or counterexample of the theory. So an observed possibility can be regarded as an established desired observational possibility.

Assuming that empirical criteria are primary, relative to their aesthetic value, they are the only relevant criteria as long as only observational and no theoretical terms are involved, for in that case there are no nonempirical (distributed) features that can have aesthetic value. In other words, nonempirical features only exist if a (relative) distinction between observational and theoretical terms can be made. I suppose that, in the present context, this distinction holds. Of course such a distinction between theoretical and observational terms leads to the distinction between an observational level of conceptual possibilities CPo and a theoretical (cum observational) level of conceptual possibilities CP = CPt. By means of this distinction a precise definition of empirical versus nonempirical features can be formulated: features of the first kind exclude possibilities on the observational level,

features of the second kind do not. Formally, e.g., for the second kind, a subset F of CP represents a nonempirical feature iff for all x in CPo there is at least one y in CP such that y has x as its 'projection' in CPo. This definition may suggest that nonempirical features of theories, in particular aesthetic ones, cannot be indicative of the empirical merits and prospects of a theory. However, by way of meta-induction they can become to be conceived as indicative in this respect. In this sense, aesthetic criteria may be seen as indirect empirical criteria, though formally quite different from the two categories of empirical criteria introduced above. From now on, we will speak only of empirical criteria (and features) in the direct sense explained above.

Truth approximation by means of empirical criteria can now be defined and founded on the basis of the following, easy to prove,

COMBINED PROJECTION & SUCCESS THEOREM

If Y is closer to T than X then Y is at least as successful as X, in the sense that:

(DF-Success:) Explanatory clause

All established observational laws explained by X are also explained by Y (or: all established desired observational features of X are also features of Y)

(DP-Success:) Instantial clause

All observed examples of X are also examples of Y "unless X is lucky" (in other words: all observed counterexamples of Y are also counterexamples of X, "unless X is lucky")

The subclause "unless X is lucky" will be clarified later. The underlying assumption for the proof of this theorem is the correctness of the empirical data, that is to say, the observed possibilities and the observational laws that are (through an inductive leap) based on them, are correct.[17]

By means of this theorem the following argument can be defended. Assume that theory Y at time t is (two-sidedly) more successful than X in the sense suggested above: not only are the two clauses (DF- and DP-Success) fulfilled, but also Y explains at least one extra observational law and X has at least one extra observed counterexample (in other words: Y has an extra observed example). This evokes the comparative success hypothesis that Y will be lastingly more successful than X. This hypothesis is a neat empirical hypothesis of a comparative nature that can be tested by deriving and testing new test implications. As soon as this hypothesis has,

in the eyes of some scientists, been sufficiently tested, the so-called rule of success can be applied, which means that they can draw the conclusion that Y will remain more successful than X. It can be proved that this is equivalent to concluding that the observational theory that follows from Y is closer to the observational truth To (the strongest true theory that can be formulated with the observational vocabulary, thus as a subset of CPo) than X. But this conclusion is in its turn a good argument for the truth approximation hypothesis (TAH) on a theoretical level: Y is closer to the (theoretical) truth T = Tt than X. In other words, the rule of success is functional for truth approximation. For this, three specific reasons can be given (for details see Kuipers 1997 or 2000, 162, 214): (1) TAH explains, according to the theorem, why Y is at least as successful as X, (2) the reversed hypothesis, X is closer to the truth than Y (RTAH), is excluded by the theorem, and (3) the denial of TAH (which is much weaker than RTAH) calls for a specific explanation of the difference in success.

5. THE IMPORTANCE OF EMPIRICAL AND AESTHETIC CONSIDERATIONS

Following the above reconstruction and justification of empirical considerations for the choice between theories, the role of nonempirical considerations can now be analysed. Although in principle all possible nonempirical considerations are involved, the analysis to be presented seems applicable especially to nonempirical aesthetic features because their formal role can be fully treated. Recall that nonempirical aesthetic features are features without empirical content that have acquired (meta-) inductive support and (hence, due to aesthetic induction) aesthetic value. By the way, it is easy to see how aesthetic value of empirical features could come into play.

Before indicating the role of aesthetic features, we must try to determine their objective importance, including their relative importance in comparison with empirical considerations. Whereas for aesthetic criteria as yet only the desired features play a part (later on I will briefly return to undesired aesthetic features), empirical criteria are concerned with established desired features, the implication of an observational law, that is, explanatory success, and established undesired features, which can be reconstructed in terms of excluding observed possibilities.

Our point of departure is the assumption that a distinction has been drawn between a theoretical and an observational level and the expectation, based on inductive grounds, that all successful theories, and hence the (relevant) truth will have a certain aesthetic feature. This feature is therefore

'held to be desired'. I also presuppose that some observed possibilities and observational laws have already been established.

Firstly I will define three kinds of similarities and differences between theories X and Y. A difference is to be understood as a difference in favour of Y, unless the contrary is stated explicitly. Later we will see that the three differences, seen as advantages of Y, come with very specific qualifications. An observed regularity that is explained by both theories X and Y or by neither will be called an E(xplanatory)-similarity; by an E-difference I will mean an observed regularity that is explained by Y but not by X. Analogously, an I(nstantial)-similarity is an observed possibility that is admitted by both or neither theory, and an I-difference is an observed possibility that is admitted by Y but not by X. Lastly, an A(esthetic)-similarity is a supposedly desired aesthetic feature that both X and Y or neither have, and an A-difference is a desired aesthetic feature that Y has, and X does not.

Especially where the differences are concerned, it is important to distinguish between possible hardness and reliability as signposts to the truth. I will argue as follows. I-differences are harder than E-differences in the sense that the correctness of I-differences is easier to determine than that of E-differences. In the same sense E-differences are harder than A-differences. But, if correct, E- and A-differences are reliable, be it modest, signposts to the truth, whereas I-differences are not.[18]

The correct determination of an I-difference requires only a correct description of an observed possibility that suffices to prove that X does not allow this possibility but Y does. The correct determination of an E-difference is less simple. It presupposes proofs to the effect that the alleged law is not explainable by means of X but can be explained by Y, and secondly that the alleged law is true (for all physical possibilities). The assumption that the alleged law is true must first of all be based on a feature that a number of observed possibilities have in common, assuming that no mistakes in the observation and measurement have been made, and secondly on 'object-(level-)induction' or, more precisely, on 'observational induction' or 'inductive generalisation' to the (time- and place-independent) general validity of that feature for all desired, that is to say, all physical possibilities. If the law concerned gets falsified after all, the E-difference is nullified. The correct determination of an A-difference, lastly, is even more difficult. It presupposes not only proofs that Y has an aesthetic feature that X does not have, but also that this feature is a feature of the truth and therefore a desired one. Generally the latter is less easy to determine than in the case of an observational feature, because the desirability of an aesthetic feature cannot be based on 'object-induction'

but at the most on (cognitive) 'meta-induction', as discussed in Section 2. It is evident that, however tempting or even plausible, this is always very risky. In summary, an E-difference is less hard than an I-difference but harder than an A-difference. Note that the differences in hardness are not so much a matter of the required proofs, but rather related to the empirical justification of the desirability of a possibility or of a (observational or aesthetic) feature.

1. *By experiment*, leading to established physical possibilities to be admitted, and opening the way for an instantial difference and hence for instantial empirical progress.

2. *By object-induction* on these experiments, leading to established observational laws to be explained, and opening the way for an explanatory difference and hence for explanatory empirical progress.

3. *By meta-induction* on empirically successful (consecutive and/or related) theories (all belonging perhaps to one research program), leading to nonempirical, notably aesthetic, features to be satisfied, and opening the way for a nonempirical, notably an aesthetic, difference and hence for nonempirical, notably aesthetic, progress.

Scheme 2. Methods of establishing desired possibilities and features.

Scheme 2 summarizes the three ways in which desired possibilities and features can be established and hence the corresponding similarities and, more importantly, differences and types of progress.

Assuming we have obtained some correct similarities and differences, what then is their relation to claims of truth approximation on a theoretical level, i.e., TAH ("Y is closer to T than X") and RTAH ("X is closer to T than Y")? For the sake of convenience I start with E-similarities and E-differences. It is evident that an E-similarity confirms the DF-clause and therefore TAH, but also the 'reversed DF-clause' and therefore RTAH. An E-difference on the other hand verifies the DF+-clause, and thus confirms TAH, but falsifies the reversed DF-clause and therefore RTAH. For these reasons I call an E-difference an indicator of presumably being closer to the truth or, simply, a signpost to the truth. The signpost metaphor is a bit misleading in that a signpost on a crossroads, if correctly situated, indicates the one out of four directions that is the right one. In this manner of speaking, an E-difference is more similar to a signpost that tells you: do not take this direction, in this case X! In a word, it is a modest indicator, due to its Popperian, negative flavour. However, in as far as it is correctly determined, an E-difference is a reliable, albeit modest, signpost to the truth, that is, a good argument in favour of TAH.

Turning to aesthetic features it is important to note first that the explanatory clause of the combined projection and success theorem can be generalized to all established desired features: if Y is closer to T than X

then all established desired (observational and non-observational) features of X are also features of Y. Hence, since a presumably desired aesthetic feature is just as much a feature of a theory as a desired observational one, the formal roles of E- and A-similarities are fully analogous. The same goes for E- and A-differences. Correct A-differences can therefore be considered, just like correct E-differences, as indicators of presumably being closer to the truth or, simply, as signposts to the truth. To recapitulate, an A-difference is (much) less hard than an E-difference. However, as far as it is correctly determined, an A-difference, too, is a reliable, modest, signpost to the truth, that is, a good argument in favour of TAH, in principle just as reliable as an E-difference.

Let us now turn to I-similarities and differences. Here the situation is more complex. Because of the 'one-many' character of the relation between the observational and theoretical levels of conceptual possibilities, a theory can have an observational feature on the observational level only if it has one on the theoretical level. The admission of an observational possibility on the theoretical level, though, cannot be based only on the admission of a suitable desired theoretical possibility, but must be based also on a suitable undesired theoretical possibility. If the observed example can be based on some admitted desired theoretical possibility, it may be called a real success of the theory. However, if the observed example can only be based on admitted undesired possibilities, it is some sort of lucky hit of that theory. For I-similarities there are all kinds of possibilities for this to occur, but it is not worth the effort of spelling them all out. I-differences, on the other hand, are very interesting. An I-difference can be based on a lucky hit of Y, in which case the DP+-clause, on the theoretical-cum-observational level, will not be verified and so TAH will not be confirmed (and therefore the reversed DP-clause and RTAH are not falsified). Of course, if it is a real success of Y, the DP+-clause is verified, TAH is confirmed, and the reversed DP-clause and RTAH are falsified. These possibilities are depicted in Figure 5.

In Figure 5, the observed example **a** is a real success of Y if there is a theoretical version in area 4 and it is a lucky hit if there is not, in which case there must be versions in 1 and 2. If (DF) [(⇔(UP)] holds, area 2 is empty, so **a** must be a real success. It is clear that whether an extra success of Y is real or only apparent cannot be ascertained on the basis of the observed example. We can say, though, that if TAH (especially (DF)) is true, the example must be real. As said, in that case the DF+-clause is verified and TAH confirmed. Although this is not a completely circular confirmation, it is a '(DF)-laden' and therefore 'TAH-laden' confirmation. So an I-difference is not reliable as a (modest) signpost, even when it

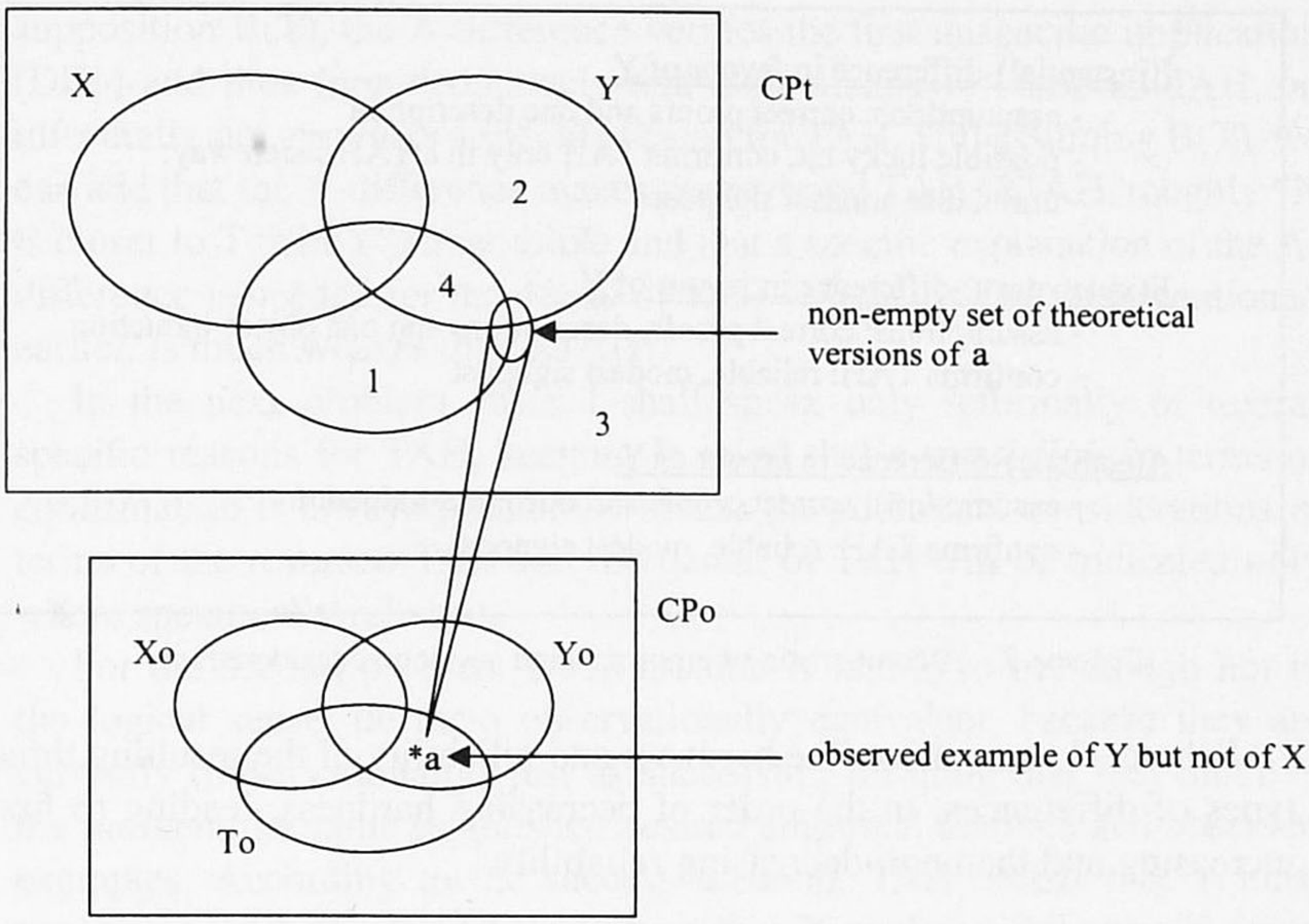


Figure 5. I-difference: observed example of Y (but not of X) as a real success or as a lucky hit. Further explanation in the text.

is correctly determined. In summary, we can say that an I-difference is, though harder than an E-difference and a lot harder than an A-difference, even when correctly determined, an unreliable (modest) signpost, for the difference can be based on a lucky hit.[19]

In total, E- and A-differences are, if correctly determined, equally good arguments in favour of TAH, but E-differences are generally much better supportable than A-differences. I-differences, on the other hand, though even better supportable than E-differences, do not constitute, if correctly determined, good arguments in favour of TAH because they are TAH-laden. I-differences are therefore difficult to weigh against E- and A-differences. Or, to return to the signpost metaphor: A-differences are more difficult to support than E-differences, but, if correctly determined, they are equally reliable signposts. I-differences, in contrast, are relatively easy to support, but, even if correctly determined, they are not reliable signposts. The consequence of all this is that, in terms of hardness and reliability, E-differences are more important than A-differences, because of the greater hardness, that is, less risky induction, and equal (positive) reliability, but the relative importance of I-differences in relation to both E- and A-differences cannot be characterised unambiguously, because they combine greater hardness, no induction being involved, with structural unreliability.

I(instantial)-difference in favour of Y
- assumptions: correct proofs and one description
- possible lucky hit, confirms TAH only in a TAH-laden way: unreliable, modest signpost

E(xplanatory)-difference in favour of Y
- assumptions: correct proofs, descriptions and one object-induction
- confirms TAH: reliable, modest signpost

A(esthetic)-difference in favour of Y
- assumptions: correct proofs and one meta-induction
- confirms TAH: reliable, modest signpost

Scheme 3. A comparison of empirical and aesthetic considerations.

Scheme 3 summarizes the hardness and reliability of the resulting three types of differences, in the order of decreasing hardness, leading to first increasing and then non-decreasing reliability.

6. THE ROLE OF AESTHETIC FEATURES

In order to describe the role of aesthetic features I will outline a number of problem cases in which the question about this role comes up. Except for the last, 'revolutionary' example, the starting-point will be that theory Y has all desired aesthetic features that X has plus one extra, which means that there is an A-difference in favour of Y. Let us denote the feature in question by B(eautiful). Since B will be counted among the desired features, the assumption is that the corresponding meta-induction is correct, that is, that T has feature B, denoted by B(T).

In the first problem case I will assume that X and Y are in a logical sense empirically, or observationally, equivalent, meaning that they have exactly the same observational consequences. The result is that they will always have equal empirical success. The claim that we can now attach to the A-difference is that it gives a specific reason for TAH. In terms of confirmation, the A-difference confirms TAH. The argument is as follows. Because of the presupposed observational equivalence of X and Y, TAH amounts, according to the equivalence theorem, to the claim that Y is at least as good when it comes to desired and undesired nonempirical features, as X (first part), and is better than X as far as some desired and undesired nonempirical features are concerned (second part). In other words, TAH implies (explains and predicts) both parts of the claim deductively, where the second part concerns two unspecific implications. Because of the pre-

supposition B(T), the A-difference verifies the first unspecific implication (DF+) and thus (non-deductively and theoretically[20]) confirms TAH, or, informally put, provides a specific reason for TAH. Still assuming B(T), we can add that the A-difference makes the reversed TAH (RTAH, roughly "X is closer to T than Y") impossible and that a specific explanation of the A-difference is needed for the denial of TAH (which, as has been mentioned earlier, is much weaker than RTAH).

In the next problem cases I shall speak only informally of (extra) specific reasons for TAH, keeping in mind that a translation in terms of confirmation is always possible, whereas the additional considerations in terms of the reversed TAH and the denial of TAH will be indicated only where specifically relevant.

For the second problem I will assume X and Y to be, though not in the logical sense, de facto observationally equivalent, because they are currently (observationally) just as successful, meaning that they share at the moment the same established desired empirical features and observed examples. According to the success theorem, TAH entails that Y must explain at least all observational laws that X explains, whereas X could have fewer counterexamples only by accident. Also following from TAH, according to the equivalence theorem, Y must be at least as good as X when it comes to nonempirical (desired and undesired) features. Another consequence of TAH is that at some time Y must be better regarding desired features, which could be empirical or nonempirical features. All three cases have to do with deductive implications, the third being of a non-specific nature. The result is that the equivalence of explanatory success can be explained by TAH, that the equivalence of instantial success can also, though with some reservation (X has no lucky hits), be explained by TAH and lastly that B can be seen, as in the first problem case, as a specific reason for TAH. The only difference is that now not only nonempirical desired features, like B, can serve as specific reasons for TAH, but also empirical features, had they been determined, which they have not. As we have seen before, the possible E-differences yet to be determined have a greater bearing than the A-difference. Hence, unlike in the first problem case, here the A-difference is a relatively weak specific reason, because stronger, empirical reasons are still feasible.

In the third problem case I take Y to have more explanatory success than X and the two theories to be instantially just as successful. The argumentation is to a high degree analogous to the previous case. The only difference is that B is now only an extra (again relatively weak) reason for TAH, on top of the extra explanatory success.

In the fourth situation I presuppose that Y is explanatorily just as successful as X but instantially more successful than X: there is at least one I-difference to Y's advantage. This situation is more complex than the last. Note first how TAH, as with the desired features, predicts only non-specifically that Y admits extra desired possibilities, but that this does not automatically imply instantial success because extra desired possibilities do not necessarily lead to extra observable possibilities. Even so, the extra instantial success of Y is surely compatible with TAH, though this is less revealing than explanatory success. More specifically formulated, only when the absence of lucky hits by Y is assumed, they exclude the reversed TAH. From now on, I will call an I-difference an instantial (empirical) reason and, for later use, an E-difference will be called an (empirical) explanatory reason. In these terms we can now say that the A-difference provides besides the instantial reasons, an extra (theoretical) reason for TAH. While an E-difference constitutes a stronger reason for TAH than an A-difference, this is not so obvious for an I-difference, as we have seen. If an I-difference were in favour not of Y but of X, it would be debatable which of the two differences, if either, would be more significant. But in the given situation, both differences point in the same direction.

From the last two situations we can easily compose the next, fifth situation: Y is explanatorily as well as instantially more successful than X. In this case, the A-difference supplies a reason for TAH additional to empirical reasons of explanatory and instantial kind.

In the sixth problem case, Y is explanatorily more successful than X but instantially not at least as successful as X. That is, against E-differences in favour of Y and no E-differences in favour of X, there are I-differences in favour of X, possibly alongside I-differences in favour of Y. As we have seen before, the I-differences in favour of X do not exclude TAH because the I-differences could be the result of lucky hits on the part of X. It is not observable whether, on the theoretical level, it is a matter of counter-examples of X, whereas it is observable that (at the observational level and therefore) on the theoretical level it is a matter of counterexamples of Y. This situation can arise because all 'X-versions' of such observed examples of X, that is, all theoretical versions belonging to X of such examples of X, may lie outside T and may therefore be undesired theoretical possibilities. As was mentioned earlier in note 19, this possibility leads to a general relativisation of additional counterexamples (of Y), especially in the case that Y is explanatorily more successful than X. An A-difference that is considered to be desired can make this relativisation more plausible. When all theoretical X-versions of all extra observable examples of X are 'outside B', they must, on the theoretical level, be counterexamples of X if

T has B, for then 'outside B' would imply 'outside T'. The result is that in this situation B can supply an extra specific reason for TAH, additional to that of the third situation. It would be interesting to check whether in the history of science we can find evidence for this theoretical possibility of using aesthetic arguments, namely as support for the relativisation of extra counterexamples of a theory that is explanatorily more successful.

Let us lastly consider a situation, the seventh, that according to McAllister is more or less typical of scientific revolutions. Let a new theory Y be explanatorily and possibly also instantially more successful than X, but without Y having all the desired aesthetic features of X. Supporters of Y, the revolutionary scientists, will of course be the first to cast doubt on the conviction that the extra aesthetic features of X should be retained. Those who cling to X, the conservatives, will remain committed to those features because they cannot imagine that the truth does not possess them. But given the weak nature of the meta-induction they may well be wrong in this, but they may also be right. However this may be, at most one of the two parties can be right. Since they assign different weights to empirical versus nonempirical criteria, they set themselves to different tasks. The challenge to the conservatives is to repair their favourite theory X, preserving its aesthetic features, in order to level the empirical scores; and they may succeed in this. The revolutionaries, who let themselves basically be guided by empirical success, have only to accommodate themselves to the adjustments of the aesthetic canon that appear to be necessary.[21]

In this way, the prevailing aesthetic criteria, the aesthetic canon, may well play a schismatic role, in the long run in favour of the revolutionaries if and only if the conservatives do not succeed in their task.

McAllister goes as far as to say that we can speak of a scientific revolution only if the theory change in question implies a rupture with the aesthetic canon. The result is, for instance, that he, in line with Kuhn, sees the transition to quantum mechanics as a revolutionary step. However, contrary to Kuhn, he also sees the transition from Copernicus's circular orbits to Kepler's elliptic orbits as a revolutionary step. Moreover, he does not see Einstein's relativity theory as a revolutionary step. Not only is there no change in the aesthetic canon, but the transition is even initiated by Einstein's attempt to apply this canon consistently. In the end this is of course a matter of definition: when does one speak of scientific revolution? In his review of McAllister's book, De Regt (1998) elaborates on this point. The only thing that is of consequence to my account is the relation between beauty and truth in the case of revolutionary transitions that include a change of the aesthetic canon. It seems that in this case at least some aesthetic induction underlying the canon and, more specific-

ally, the corresponding cognitive meta-induction was not correct, for on further investigation it turns out that the (relevant) truth does not satisfy the induced aesthetic expectations.

To be sure, 'the truth' changes in all revolutionary transitions as soon as the (theoretical) vocabulary changes. The analysis in Sections 3 and 4 essentially implies that vocabulary change does not prevent the relevance of truth approximation considerations. For example, the most extreme case is that the old theoretical vocabulary is no longer supposed to have referring terms at all, in which case its theoretical truth coincides as a matter of fact with its observational truth. The new theoretical vocabulary introduces in fact a new hypothesis about a theoretical truth that goes beyond, but is relevant for, the common observational truth, corresponding to the common observational vocabulary as far as that is not disputed (see Kuipers (2000, Ch. 9)). The foregoing discussion about 'beauty and revolution' shows in addition that aesthetic considerations can play a role in revolutionary disputes which is relevant from the truth approximation perspective.

7. SUGGESTIONS FOR FURTHER RESEARCH

Various matters await further research. I shall first touch on some possible experiments. Then I will suggest some formal elaborations. Finally, I will suggest some possible connections with other perspectives and problems that seem worth investigating.

In Section 2 I have argued that the aesthetic induction may be a variant of the mere-exposure effect, viz., the qualified-exposure effect. The naturalized approach suggests several experiments with normal and toy pieces of art and with scientific examples to establish the conditions and limitations of the effect. In particular, the presumably strengthening and lengthening role of variation on a theme, comparable to theory revisions, and, similarly, the role of a counterpart to an empirical success, as a kind of extra reward, seem worth investigating. For the first, one might think of Bach's successive two-part and three-part Inventionen, for the second of some kind of utility, as in McAllister's case of cast-iron architecture. Moreover, further evidence for the varying character of the aesthetic canon when different phases or different research programmes of the same discipline and different disciplines are compared would strengthen the basic ideas around aesthetic induction as such and its diagnosis as a variant of the mere-exposure effect. Finally, my refined claim about aesthetic induction can be falsified: determine a nonempirical, (not necessarily) distributed feature which happens to accompany all increasingly successful theories in

a certain area from a certain stage on and which is not generally considered beautiful, and increasingly so, by the relevant scientists.

The formal analysis that followed after Section 2 was based on conceiving the beauty of theories and of the truth in terms of their nonempirical features, and they were exclusively interpreted in a distributed way, namely as shared properties of all possibilities admitted by them. Although I have argued that this leaves room for many more features than one might think at first sight, it is worth to look for examples of features that cannot be reconstructed in this way. One type might be so-called constraints. According to the structuralist view of theories, in terms of which the formal analysis in this article is presented, theories are not only seen as sets of conceptual possibilities, but many additional restrictions are involved, notably so-called constraints. For instance, an identity constraint says that a certain function, such as the mass function, must in every application assign the same (rest) mass to the same object. Constraints can also have acquired inductive support and aesthetic value and can be represented as a certain kind of set of sets of conceptual possibilities. It seems perfectly possible to extend the basic truth approximation theory for this kind of restrictions in general and for their aesthetic value in particular. The reason is that constraints have the so-called 'subset property': if a set of possibilities satisfies a constraint, all its subsets do so, including the singleton sets. Thus, constraints are distributed features of sorts.

Of course, there may well be aesthetic features that can neither be represented as a set of conceptual possibilities nor as a set of such sets. They may be of a more holistic kind. For example, a theory may be called symmetric not only because of its symmetric possibilities, but also because it is closed under a certain operation: given a model, applying the operation leads again to a model of the theory. Other examples of holistic, at least non-distributed, features of theories are diversity (of admitted/desired possibilities) and convexity. In general, all formal features that postulate membership claims in response to given members cannot be distributed. For such non-distributed features an alternative formal analysis will have to be found to complete the naturalistic analysis to a full-fledged naturalistic-cum-formal analysis of such features.

One plausible option is immediately suggested by the strong, but nonetheless partial, analogy between truth-oriented research (of a nomological nature) and 'design research', as elaborated in (Kuipers et al. 1992). Design research concentrates on making or improving certain products or processes. The equivalence theorem not only makes possible a better formulation of this analogy, but also strengthens the analogy because the logic of design research is defined, as a matter of course, in terms of fea-

tures of the target product and of a realised prototype. If only distributed features are taken into account, truth approximation can be redescribed as an attempt to improve a theory by enlarging its set of desired features and diminishing its set of undesired features, where the (un-)desired features are based on the (unknown) strongest true theory, as defined in Section 3. This suggests defining truthlikeness in this way without the restriction to distributed features, which would of course imply the restricted re-described one, and hence the original 'possibility definition'. As far as the naturalistic part of our analysis is concerned such an unrestricted feature definition is unproblematic. However, the formal part of the analysis would lead to a problem. Not only 'having a true consequence' would belong, for all true consequences, to the desired features, but also 'not having a false consequence' would belong, for all false consequences, to the undesired features, for by definition 'the truth' does not have any false consequence. Note first that, whereas the former type of feature is distributed, the latter is not. Moreover, as we have indicated in note 14, Popper's original definition failed precisely because it was phrased in terms of (all) true and false consequences: it left no logical room for some false theories, being closer to the truth than other false theories. Hence, a definition of truthlikeness in terms of features needs some restriction of the type of features in order to avoid this inadequacy. Thus, the remaining question is for what types of features, other than distributed features, a feature definition of truthlikeness is adequate in this sense or even equivalent to the possibility definition. Of course, it may well be that there are such types of features that include the suggested holistic features representing, for example, (a kind of) symmetry and diversity.

The analysis was further restricted to (distributed) desired aesthetic features, but of course, at first sight, there are also *un*desired aesthetic features: features that we find ugly and strive to avoid. The question is whether those can be accounted for by the (UF)-clause in accordance with the informal use. This seems plausible as far as negative aesthetic appreciation for a nonempirical feature can be generated in a similar systematic way as its positive counterpart by aesthetic induction. But this is not obvious, as far as such a feature is not the formal counterpart, in one way or another, of a positively appreciated feature, and if this is the case separate treatment seems redundant.

In the truth approximation theory the reference of theoretical terms and the idea of a better approach of 'the referential truth' (Kuipers 2000, Ch. 9) can be expressed as well. Here, too, aesthetic considerations can be involved, for instance in the form of ontological economy considerations (Occam's razor) as a well-known form of simplicity. Within this frame-

work it is relevant to note that there appears to be a connection between the equivalence theorem of Section 3 and (the strong version of) Leibniz's theory of identity (two objects are identical if (and only if) they have the same features) applied to sets.

There is also a 'refined' version of the theory, according to which not all admitted undesired possibilities are equally unfavourable, with the consequence that improvement of theories is possible by replacing undesired possibilities by less unfavourable, but formally still undesired, ones. Since, according to the equivalence theorem, undesired possibilities have to do with desired features, this also means that a refinement of aesthetic features and their role seems possible.

The current treatment of aesthetic features is strongly determined by the qualitative nature of the truth approximation theory used. It is difficult to imagine how the quantitative theory of Niiniluoto (1987) could be employed for aesthetic features, but the present article certainly offers a challenge to do so. However this may be, the application of the quantitative theory is much more limited, namely to those areas where a meaningful function of the distance between the relevant conceptual possibilities can be defined. But when this can be done, a quantitative definition results in a linear ordering of theories whereas a qualitative definition leads only to a partial ordering.

According to Thagard (1988, 1992) simplicity can compensate for empirical shortcomings as a result of which on closer inspection it becomes possible that a theory is chosen that is more simple but empirically less successful. In the light of the above analysis, this does not seem justifiable for explanatory success, but because lucky hits are possible, it may be reasonable for instantial success. Nevertheless, it seems difficult to formalise this within a qualitative approach, but it should be possible within a quantitative approach.

So far a number of suggestions for further formal elaboration of this article. But other questions crop up too. Laudan (1977) distinguishes between empirical and conceptual problems. It seems reasonable to take empirical problems of a theory as observed counterexamples and unexplained laws. The suggestion of this article is that conceptual problems can be taken as a lack of certain desired nonempirical features or perhaps as the presence of certain undesired nonempirical features. Especially for the former interpretation it is obvious that the role of conceptual problems will then be formally analogous to the role of aesthetic problems, that is to say the lack of certain (desired) aesthetic features. Analogously the formal role of a conceptual problem that is solved by a new theory can be compared with having a certain desired aesthetic feature.

Besides the partial analogy mentioned above between truth-oriented research and design research, another basic form of scientific research, viz., explication of concepts, turns out also to be partially analogous to truth approximation research. Obvious examples and non-examples of an intuitive concept function as (fixed) desired, and undesired possibilities, respectively, whereas conditions of adequacy can be seen as (fixed) desired features of the intended concept. Thus a 'general logic of research', with different specifications, appears possible (Kuipers 2001, Ch. 9). In view of the present paper, all these types of research leave room for a modest role of aesthetic considerations.

Lastly the connection between 'simplicity', 'expected predictive success', and 'truth approximation' must be examined for so-called curve fitting. In the literature on simplicity for curve fitting (Sober 1998) the Akaike's theorem plays a central role. It specifies an unbiased estimate of the predictive accuracy of a family of (e.g., linear or parabolic) curves, also called its closeness to the truth, based on past performance. The simplicity or, rather, the complexity of a family is measured in terms of the number of parameters that is characteristic of that family. As far as complexity is concerned, Akaike's theorem states that the estimated predictive accuracy decreases with increasing complexity. Though the breadth of this theorem is heavily criticised (Kieseppä 1997) and the type of simplicity is not a straightforward nonempirical distributed feature of families of curves, it seems logical to wonder what its relation is to truth approximation in the sense of this paper.

8. CONCLUSION

Even without further investigations we can draw the following conclusions from the presented naturalistic-cum-formal analysis. First of all, as could be expected, scientists' intuitions are correct. Aesthetic considerations can usefully be put into service, even though they are less hard than empirical considerations of an explanatory nature. They can nevertheless function as modest signposts to the truth. Secondly, McAllister could very well be right concerning the origin of aesthetic considerations in general: they may arise from aesthetic induction. This was argued to be partly a matter of cognitive meta-induction and partly a variant of the mere-exposure effect, called affective induction. As for their role in scientific revolutions in particular, aesthetic considerations can be far from the truth and therefore obstruct scientific progress among aesthetically conservative scientists.

The core of the answer to the title question "Beauty, a road to the truth?" hence is: yes, beauty can be a road to the truth as far as the truth is

beautiful in the specific sense that it has distributed features that we have come to experience as beautiful. This is a nontrivial answer because it is not immediately obvious that a common feature of a theory and the truth can be considered as a (modest) signpost to the truth. The answer is also a demystifying and perhaps disenchanting one because the analysis does not specifically relate to the (acquired) aesthetic value of the feature, but to its formal aspect, namely which conceptual possibilities it includes and excludes and to its meta-inductive support, that is, the legitimate inductive reasons there are to assume that the truth has this formal feature. As a consequence, every nonempirical feature can indicate the road to the truth if there are inductive grounds for assuming that the truth probably has this feature. Contrary to Dirac and Einstein (McAllister 1996, Ch. 6) one does not need to assume that the connection between truth and beauty is intrinsic. As for every form of cognitive induction, it is sufficient that the meta-induction involved in aesthetic induction is on average justified at least as often as it is beside the mark.[22]

The outline of the seven problem cases shows that the answer is also useful. Depending on the situation, an aesthetic feature that is assumed to be desired can play a specific role. These features must be handled prudently, though. First, one must keep a close watch over the necessary relativisation of the feature in terms of the meta-induction. Second, an aesthetic feature can, for the same reason, hardly compete with explanatory success that points in a different direction. So one should realise that one might retard some revolutionary change by sticking to them. But, as we have seen in the sixth problem case, an aesthetic feature can very well raise an objection against instantial success that points in a different direction.

Put differently, though there is a clear difference in weight between desired empirical and nonempirical features, the case is different when it comes to the relative importance of differences in instantial success. If the relevant data and the accompanying inductive generalisation are correct, the truth has the empirical feature in question. Though only correct data are needed for a difference in instantial success, such a difference is not a reliable signpost to the truth when a distinction between theoretical and observational terms is operative. As for aesthetic and, more generally, nonempirical features, presupposing them remains a meta-inductive gamble waiting for decisive empirical arguments. That is, subsequent empirical success may make them superfluous or they may downplay their weight. Such new empirical arguments may or may not be the result of a well motivated change in the difference between theoretical and observational terms. The standard form of such a change is the transformation of theoretical terms into observational ones, by accepting certain theories as

true. In this way the former theoretical terms become observable or at least measurable.[23]

NOTES

[1] This paper is an enlarged and revised version of a translation by Eefke Meijer of my Dutch paper: "Kan schoonheid de weg wijzen naar de waarheid?", *Algemeen Nederlands Tijdschrift voor Wijsbegeerte* (ANTW), 91.3, 1999, 174–193. James McAllister corrected the English. The paper was read in 1998 at a meeting of the General Dutch Union for Philosophy (ANVW), in Utrecht, in 1999 at the University of Trieste, and in 2000 at a meeting of the Forum for European Culture, in Amsterdam, at the Catholic University of Lublin, at the annual meeting of the British Society for the Philosophy of Science (BSPS) in Sheffield, and at the Gesellschaft für Analytische Philosophie (GAP-4) congress in Bielefeld. I thank David Atkinson, Joop Doorman, Job van Eck, Roberto Festa, Erik Krabbe, Jan Albert van Laar, Anne Ruth Mackor, Jeanne Peijnenburg, Henk de Regt and André de Vries for comments on the Dutch version and Dirk Povel for his information leading to the literature on the mere-exposure effect. I thank David Miller in particular for his extensive and critical comments on the first English translation, which he presented at the BSPS meeting in Sheffield. I also thank Jeffrey Koperski and Elliott Sober for their suggestions on the same occasion. Finally, I like to thank the three anonymous referees who pressed me to clarify a number points.

[2] The 25 interviews are available on videotape (http://www.vpro.nl/frontend/index.shtml) and have also been published in Dutch under the title *Het boek over de schoonheid en de troost (The Book Concerning Beauty and Consolation)* (Kayzer, 2000). Unlike many of the other interviews, those with Gould and Weinberg have been transcribed and translated fairly literally.

[3] E.g., many of those attending my presentation on the BSPS meeting in Sheffield in July 2000.

[4] As a matter of fact, in science we come across two kinds of beauty. We speak of the beauty of methods of proof and problem solving on the one hand, and of results such as propositions, laws, theories, and truths on the other. The so-called diagonal proof of the non-denumerability of real numbers is an example of a method that strikes almost everyone for its simplicity and inventiveness. In Kuipers (1991) I have collected ten examples of beautiful problem-solving methods for quite mundane problems such as "What is the shortest network of roads between four cities located on the corners of an imaginary square?". However, as suggested, and as in this case, solutions themselves may also be considered as beautiful, like new results in general. Moreover, regarding results themselves, we might distinguish between new results that are found beautiful because they are surprising, perhaps by opening new perspectives, and results that are found beautiful because they fit into the current 'aesthetic canon'. This paper addresses the last type of aesthetic considerations. I thank Michael Stölzner for pressing me to make the latter distinction explicit.

[5] This is not to suggest that standard examples of aesthetic features mentioned by physicists do not play a role in biology. For example, Gould mentions order a number of times and, as Sober (2000) points out, simplicity in the form of parsimony plays a considerable role in taxonomy.

[6] For the gravedigger example and some other ones, illustrating the same point, see also Weinberg (1993, 119).

[7] McAllister (1996) deals in particular with symmetry, simplicity and visualizability, and their opposites, see below. Weinberg (1993) deals in chapter 6, entitled "Beautiful theories", not only with inevitability or rigidity, but also with simplicity and symmetry.

[8] Our terminology of cognitive, affective and (later) behavioural induction is inspired by Ye's (2000) distinction of cognitive, affective and behavioural priming.

[9] A telling example of this is the standard model of elementary particles, combining the electroweak theory with quantum chromodynamics. It is not uniquely determined from symmetry considerations alone, but it is further restricted in form by the requirement that certain infinities cancel in the calculation of physical quantities (the 'renormalizability condition' of nonabelian gauge theories applied by – 1999 Nobel Prize winners – Gerard 't Hooft and Martinus Veltman (see e.g., Weinberg (1993, 95–96, 117)).

[10] For the most recent and complete version of the basic (and refined) theory on truth-likeness and truth approximation, see (Kuipers 2000). For a concise version, see (Kuipers 1997).

[11] The following standard set-theoretical notation will be used. For sets A and B, the intersection will be designated by 'A ∩ B', which means the set of all common elements; 'A-B' indicates the difference, which is the set of elements that are elements of A but not elements of B. 'A ⊆ B' indicates that A is a *subset* of B and 'A ⊂ B' that A is a *proper subset* of B. The first case does not exclude 'A = B', the second does. 'Ø' indicates the *empty* set.

[12] In the second definition 'two-sided' refers to the fact that two clauses are postulated, while one-sided closer to the truth postulates only one of them. For completeness I give a set-theoretical translation of the clauses:

(DP)	$X \cap T \subseteq Y \cap T$	(DP+)	$(Y \cap T) - X \neq \emptyset$
(UP)	$Y - T \subseteq X - T$	(UP+)	$(X - T) - Y \neq \emptyset$

[13] However, it is easy to check that 'being true' of a theory X in the weak sense that T is a subset of X, is not a distributed feature, let alone 'being true' in the strong sense of the claim "T = X".

[14] Popper has given a definition of 'closer to the truth' in terms of more true and fewer false consequences, that was acknowledged later (also by Popper himself) to be unsound. In terms of features Popper's mistake can be rephrased as an exceedingly broad understanding of undesired features: not only the features that are defined undesired in the above, but also the neutral features fall under Popper's definition. For further analysis, see Kuipers (1997, 2000, Section 8.1) and also (Zwart 1998, Chapter 2), who has creatively reused part of Popper's intuitions (Chapter 6).

[15] The set-theoretical interpretation of the universe of features and of the common element and the set-theoretical characterisation of the new clauses can simply be given in terms of 'powersets' and 'co-powersets'. The powerset P(X) of X is defined as the set of all subsets of X. The rectangle representing the 'universe' of all possibly relevant, distributed, features, can now be interpreted as the 'powerset' P(CP) of CP. Like a kind of mirror notion to that of powerset, the co-powerset Q(X) of X is the set of all subsets of CP that include X, also called the supersets of X (within CP). Q(X) then represents the features of X, Q(T) the desired features and Q(CP-T) the undesired features. Note that Q(T) and Q(CP-T) have exactly one set as common element, namely CP, that corresponds with the tautology, and

that is of course included in the set of features of every theory. This results in the following formal translations of the four feature clauses:

$$\text{(UF)}\quad Q(Y) \cap Q(\text{CP-T}) \subseteq Q(X) \cap Q(\text{CP-T}) \qquad \text{(UF+)}\quad (Q(X) \cap Q(\text{CP-T})) - Q(Y) \neq \emptyset$$
$$\text{(DF)}\quad Q(X) \cap Q(T) \subseteq Q(Y) \cap Q(T) \qquad \text{(DF+)}\quad (Q(Y) \cap Q(T)) - Q(X) \neq \emptyset$$

Proving the equivalence theses in terms of sets now becomes a nice exercise in 'set calculation'.

[16] Let us give, by way of example, a proof of the claim that (UF) entails (DP). Assume (UF) and let, contrary to (DP), x be a desired possibility admitted by X, that is, x belongs to $X \cap T$, and let x not be admitted by Y, hence belong to T-Y. Now CP-{x} is a superset of Y, hence it represents a feature of Y which only excludes desired possibilities, viz. x, (and no undesired ones). Hence it is an undesired feature of Y, which should according to (UF) also be a feature of X, excluding that x is a member of X. Q.e.d. All proofs are of this elementary nature.

[17] For the set-theoretical formulation of this theorem I refer to (Kuipers 2000, Sections 7.3.3 and 9.1.1).

[18] The hardness of differences will be defined in relation to the kind of inductive generalisation involved, whereas their reliability as signposts to the truth will be interpreted in relation to the transition from the observational level to the theoretical one.

[19] This nature of I-differences makes it possible for realists who want to defend TAH in a concrete case, to relativise reversed I-differences: after all, an instantial success of X that is a counterexample of Y could be a lucky hit on the part of X.

[20] The confirmation of TAH is called 'non-deductive' because the specific A-difference does not follow deductively from TAH, though it does make TAH more plausible. See (Kuipers 2000, Part I) for a coherent analysis of deductive, non-deductive, and inductive confirmation. The confirmation is called 'theoretical' because B is a nonempirical feature.

[21] The temporal opposite of the seventh situation is also interesting. Assume that a new theory Y is very beautiful, but has extra empirical problems compared to X. Then one can defend the position that the problems do not actually arise from Y, but from additional presuppositions that need to be made. Dirac's thought experiment concerning negative results for general relativity theory (McAllister 1996, 93–94) is an example of this.

[22] Contrary to the relativisation that Derksen (1999) attempts to perform on McAllister's inductive argument for aesthetic features in a narrower sense, that argument appears to have nothing to do with the specific aesthetic value of aesthetic features, but exclusively with the formal and meta-inductive aspects of such features and of other nonempirical features.

[23] For this long-term dynamics in science, see Kuipers (2000, Sections 9.2 and 13.3).

REFERENCES

Bornstein, R. F.: 1989, 'Exposure and Affect: Overview and Meta-Analysis of Research, 1968–1987', *Psychological Bulletin* **106**(2), 265–289.

Bornstein, R. F.: 1994, 'Are Subliminal Mere Exposure Effects a Form of Implicit Learning?', *Behavioral and Brain Sciences* **17**, 398–399.

Derksen, T.: 1999, 'Schoonheid als argument. Over James W. McAllister, Beauty and Revolution in Science', *Algemeen Nederlands Tijdschrift voor Wijsbegeerte* **91**(3), 168–173.

Kayzer, W.: 2000, *Het boek over de schoonheid en de troost*, Contact, Amsterdam.
Kieseppä, I.: 1997, 'Akaike Information Criterion Curve-Fitting, and the Philosophical Problem of Simplicity', *The British Journal for the Philosophy of Science* **48**(1), 21–48.
Kuipers, T.: 1991, 'Dat vind ik nou mooi', in R. Segers (ed.), *Visies op cultuur en literatuur. Opstellen naar aanleiding van het werk van J.J.A. Mooij*, Rodopi, Amsterdam, pp. 69–75.
Kuipers, T.: 1997, 'The Dual Foundation of Qualitative Truth Approximation', *Erkenntnis* **47**(2), 145–179.
Kuipers, T.: 2000, *From Instrumentalism to Constructive Realism*, Synthese Library, Vol. 287, Kluwer Academic Publishers, Dordrecht.
Kuipers, T.: 2001, *Structures in Science*, Synthese Library, Vol. 301, Kluwer Academic Publishers, Dordrecht.
Kuipers, T., R. Vos, and H. Sie: 1992, 'Design Research Programs and Logic of Their Development', *Erkenntnis* **37**(1), 37–63.
Laudan, L.: 1977, *Progress and its Problems*, University of California Press, Berkeley, CA.
McAllister, J.: 1996, *Beauty and Revolution in Science*, Cornell University Press, Ithaca, NY.
McAllister, J.: 1998, 'Is Beauty a Sign of Truth in Scientific Theories?', *American Scientist* **86**, 174–183.
McAllister, J.: 1999, 'Waarheid en schoonheid in de wetenschap', *Algemeen Nederlands Tijdschrift voor Wijsbegeerte* **93**(1), 153–167.
Mull, H. K.: 1957, 'The Effect of Repetition Upon the Enjoyment of Modern Music', *The Journal of Psychology* **43**, 155–162.
Niiniluoto, I.: 1987, *Truthlikeness*, Reidel, Dordrecht.
Regt, Henk de: 1998, 'Explaining the Splendour of Science', *Studies in History and Philosophy of Science* **29**(1), 155–165.
Seamon, J. G. et al.: 1995, 'The Mere Exposure Effect is Based on Implicit Memory', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**(3), 711–721.
Sober, E.: 1998, 'Simplicity (in Scientific Theories)', *Routledge Encyclopedia of Philosophy* **8**, 780–783.
Sober, E.: 2000[2], *Philosophy of Biology*, Oxford University Press, Oxford.
Temme, J.: 1983, *Over smaak valt te twisten. Sociaal-psychologische beïnvloedingsprocessen van esthetische waardering (Accounting for Tastes. Social Psychological Influence Processes on Aesthetic Appreciation)*. With a summary in English. Dissertation, University of Utrecht.
Thagard, P.: 1988, *Computational Philosophy of Science*, MIT Press, Cambridge, MA.
Thagard, P. (1992) *Conceptual Revolutions*, Princeton University Press, Princeton, NJ.
Vos, M. de: 1999, 'Diep in het geheim gestoken. Over duistere poëzie', in *Mooi*, special issue *De Gids* **3/4**, 166–172.
Weinberg, S.: 1993, *Dreams of a Final Theory*, Vintage, London.
Ye, G.: 2000, *Modeling the Unconscious Components of Marketing Communication: Familiarity, Decision Criteria, Primary Affect, and Mere-Exposure Effect*, Dissertation, Tilburg.
Ye, G. and W. F. Van Raaij: 1997, 'What Inhibits the Mere-Exposure Effect: Recollection or Familiarity?', *Journal of Economic Psychology* **18**, 629–648.
Zajonc, R.: 1968, *Attitudinal Effects of Mere Exposure*. Monograph supplement 9 of *The Journal of Personality and Social Psychology*.

Zwart, S.: 1998), *Approach to the Truth. Verisimilitude and Truthlikeness*, Dissertation, University of Groningen, ILLC Dissertation Series, Amsterdam. (A revised version will appear in the Synthese Library of Kluwer Academic Publishers.)

Department of Theoretical Philosophy
University of Groningen
Aweg 30
9718 CW Groningen
The Netherlands
E-mail: T.A.F.Kuipers@philos.rug.nl
www.philos.rug.nl